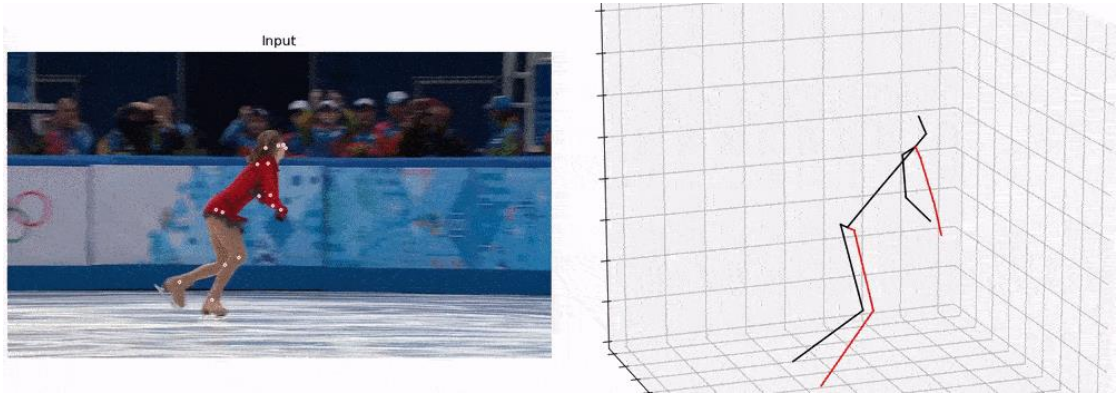
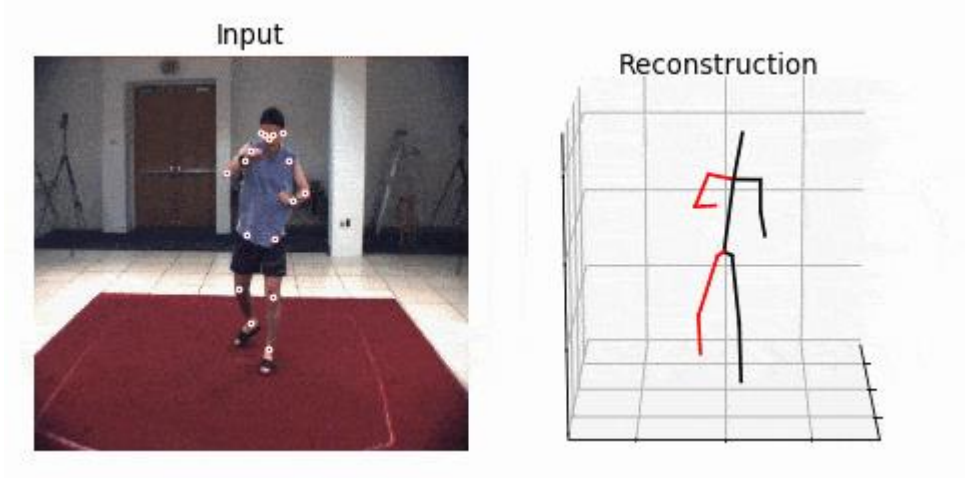
Unsupervised 3D Human Pose Estimation in Multi-view-multi-pose Video

CHENG SUN, DIEGO THOMAS, HIROSHI KAWASAKI

KYUSHU UNIVERSITY

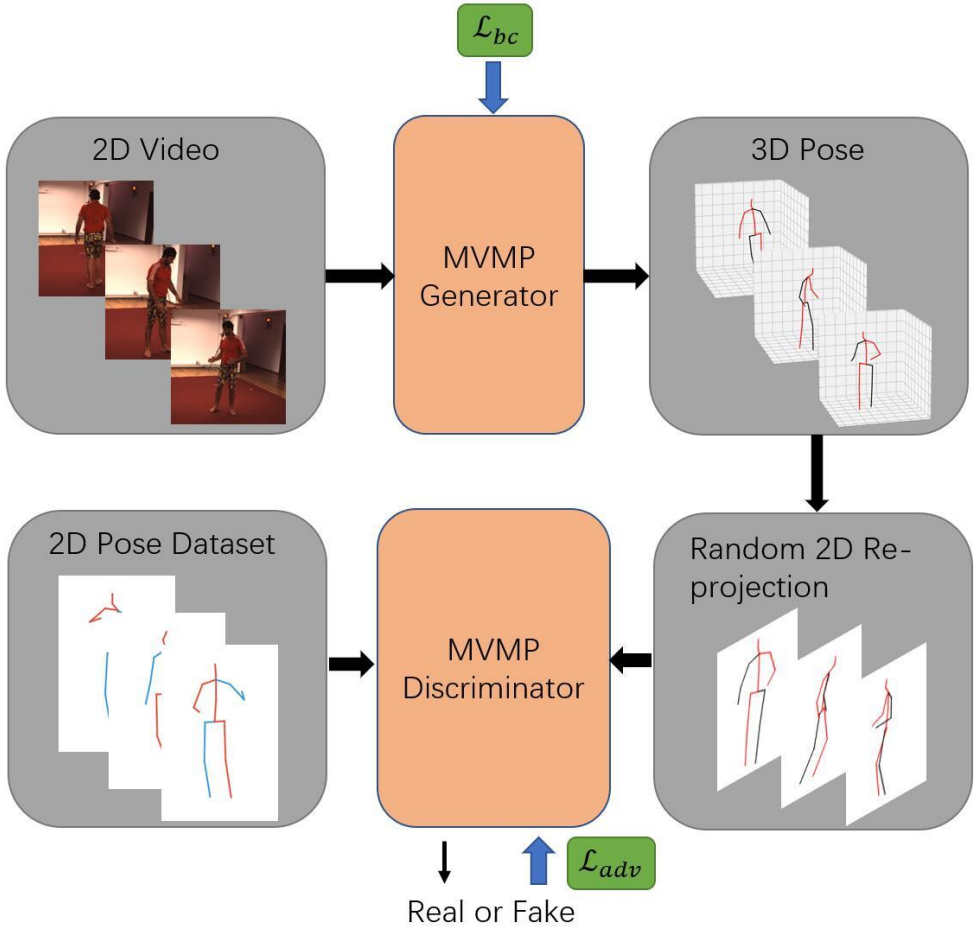
3D Human Pose Estimation



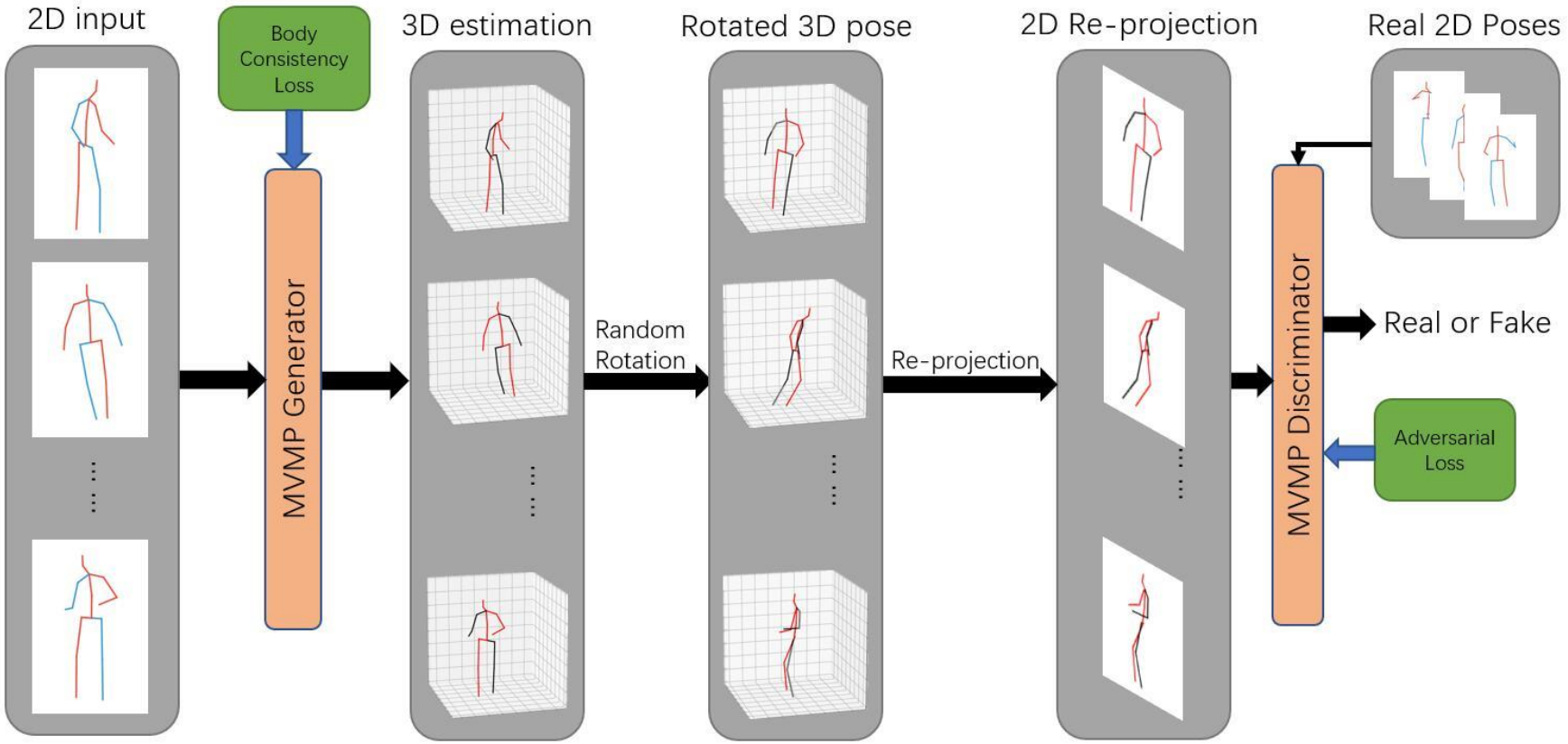


[1] Pavllo, Dario, et al. "3d human pose estimation in video with temporal convolutions and semi-supervised training." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019.

Overview


$\mathcal{L}_{bc}$
2D Video
MVMP Generator
3D Pose
2D Pose Dataset
MVMP Discriminator
Random 2D Re-projection
$\mathcal{L}_{adv}$
Real or Fake

Structure


2D input
Body Consistency Loss
MVMP Generator
3D estimation
Random Rotation
Rotated 3D pose
Re-projection
2D Re-projection
Real 2D Poses
MVMP Discriminator
Real or Fake
Adversarial Loss

Adversarial Loss

Original Adversarial Loss:

$$\mathcal{L}_{adv} = E_{x \sim P_{data}}[\log D(x)] + E_{x \sim P_G}[\log(1 - D(x))]$$

[2] Goodfellow, Ian, et al. "Generative adversarial nets." *Advances in neural information processing systems*. 2014.

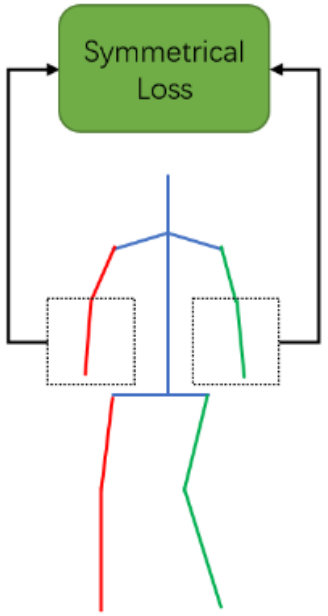
Adversarial Loss

Wasserstein Adversarial Loss:

$$\mathcal{L}_{adv} = E_{x \sim P_{data}} [f_w(x)] - E_{x \sim P_G} [f_w(x)]$$

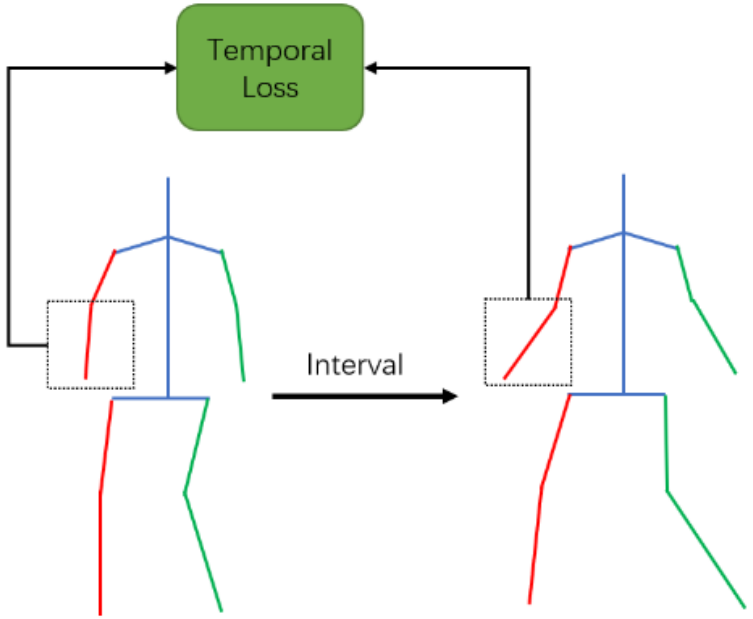
[3] Arjovsky, Martin, Soumith Chintala, and Léon Bottou. "Wasserstein gan." *arXiv preprint arXiv:1701.07875* (2017).

Body Consistency Loss



$$\mathcal{L}_{sym} = \sum_{i=1}^{k} \sum_{j \in S_{sym}} \left\| B_{ij} - B'_{ij} \right\|_2^2$$

Body Consistency Loss



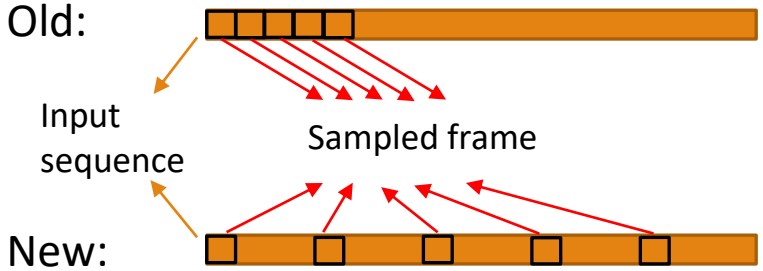
$$\mathcal{L}_{tem} = \sum_{i=1}^{k-1} \sum_{j \in S_{tem}} \left\| B_{ij} - B_{(i+1)j} \right\|_2^2$$

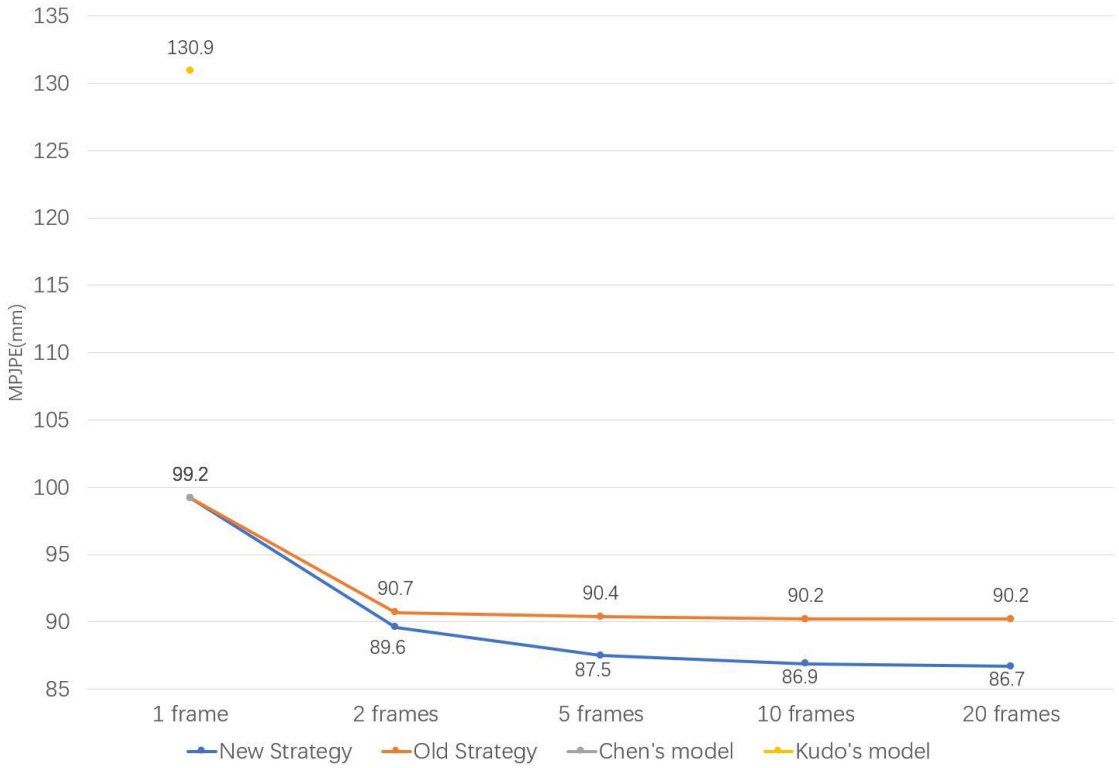
Quantitative result

Method	Dir.	Disc.	Eat	Greet	Phone	Photo	Pose	Purch.	Sit	SitD.	Smoke	Wait	WalkD.	Walk	WalkT.	Avg
supervised																
Martinez *et al.* [1]	51.8	56.2	58.1	59.0	69.5	78.4	55.2	58.1	74.0	94.6	62.3	59.1	65.1	49.5	52.4	62.9
Iskakov *et al.* [28]	19.9	20.0	18.9	18.5	20.5	19.4	18.4	22.1	22.5	28.7	21.2	20.8	19.7	22.1	20.2	20.8
self-supervised																
Kocabas *et al.* [29]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	60.6
semi-supervised																
Pavllo *et al.* [3]	45.2	46.7	43.3	45.6	48.1	55.1	44.6	44.3	57.3	65.8	47.1	44.0	49.0	32.8	33.9	46.8
weakly-supervised																
Wandt *et al.* [30]	77.5	85.2	82.7	93.8	93.9	101.0	82.9	102.6	100.5	125.8	88.0	84.8	72.6	78.8	79.0	89.9
Yang *et al.* [2]	51.5	58.9	50.4	57.0	62.1	65.4	49.8	52.7	69.2	85.2	57.4	58.4	43.6	60.1	47.7	58.6
unsupervised																
Kudo *et al.* [4]	125.0	137.9	107.2	130.8	115.1	127.3	147.7	128.7	134.7	139.8	114.5	147.1	130.8	125.6	151.1	130.9
Chen *et al.* [5]	97.1	99.4	83.2	93.8	100.3	115.4	95.2	96.9	111.4	112.7	94.1	104.1	101.5	86.3	96.5	99.2
Ours(2-frame)	89.9	92.4	78.5	91.8	93.0	97.1	88.7	86.4	97.1	101.0	89.2	98.3	90.3	71.5	79.0	89.6
Ours(5-frame)	88.0	89.6	75.0	91.3	90.9	93.5	86.8	81.5	93.7	100.3	88.0	97.2	87.6	70.8	78.4	87.5
Ours(10-frame)	87.4	89.1	74.7	90.2	90.5	93.3	86.2	80.1	93.5	99.7	87.8	96.4	87.2	70.3	77.6	86.9
Ours(20-frame)	**87.1**	**88.7**	**74.6**	**90.0**	**90.3**	**93.4**	**85.7**	**80.2**	**93.1**	**99.9**	**87.4**	**96.3**	**87.2**	**70.0**	**77.1**	**86.7**

Quantitative result

Different sampling strategy:





Qualitative result